Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Based on the analysis on the dataset the Following fields are categorical variables.

   a. Season
   b. Year
   c. Month
   d. Weekday
   e. Weather

The above variable are having their own efforts to increase the demand on the sales count for the bike sharing

2. Why is it important to use drop_first=True during dummy variable creation?

   The Drop_first=True helps in reducing the extra column created during dummy variable creation which reduce the correlation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   In the pair-plot the fields temp and atemp with correlation 0.99.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumption in linear regression using the Durbin-Watson method which is h aving 1.986 on the training data set which is a positive correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explain ing the demand of the shared bikes?

   1.atemp
   2.winter(season)
   3. Month(jul,sep,oct)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

The Linear regression algorithm is a machine learning model in which performs t he regression task and also helps in predicting the target values on the independent variable. It com es with the formula

$$Y = mx + c$$

m- slope y is the intercept of c and c is the constant

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet has four data set which are nearly identical in simple descri ptive statistics , it looks different when we normally distributed and it looks very different when we p lot the data point in the scatter box.

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a l inear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable ch anges in the same direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scali ng and standardized scaling?

Scaling is a step of data Pre-Processing which is applied to independent variable to normalize the data within a particular range. It also helps in speeding up the calculations in an alg orithm.

Normalized scaling	Standardised scaling
Min & Max feature used for scaling	Mean & standard deviation used for scaling
Feature are of different scales	Ensure zero mean and unit standard deviation
Scales values between [0,1] or [-1,1]	Scales doesn't have values
Affected by the outliers	Doesn't affected by the outliers
Its is called as scaling normalization	Its is called as z-score normalization

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

It means the perfect correlation, then VIF = infinity.
Formula for VIF =1/(1-R^2).here if R^2 values is 1 then the 1/0 will be infinity

Its happen due to the multicollinearity in one of the variable which we can drop to avoid it.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other . A quantile is a fraction where certain values fall below that quantile. For example, the median is a q uantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.